(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 August 2002 (29.08.2002)

PCT

(10) International Publication Number
WO 02/067245 A1

(51) International Patent Classification: G10L 17/00

(21) International Application Number: PCT/GB02/00665

(22) International Filing Date: 15 February 2002 (15.02.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

0103875.1	16 February 2001 (16.02.2001)	GB
0108473.0	4 April 2001 (04.04.2001)	GB

(71) Applicant *(for all designated States except US)*: **IMAGINATION TECHNOLOGIES LIMITED** [GB/GB]; Home Park Estate, Kings Langley, Hertfordshire WD4 8LZ (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **CAREY, Michael, John** [GB/GB]; c/o Imagination Technologies Limited, Turing House, Station Road, Chepstow NP6 5PB (GB). **AUCKENTHALER, Roland** [AT/GB]; 14 Wessex Street, Cardiff CF5 1LA (GB).

(74) **Agent: ROBSON, Aidan, John**; Reddie & Grose, 16 Theobalds Road, London WC1X 8PL (GB).

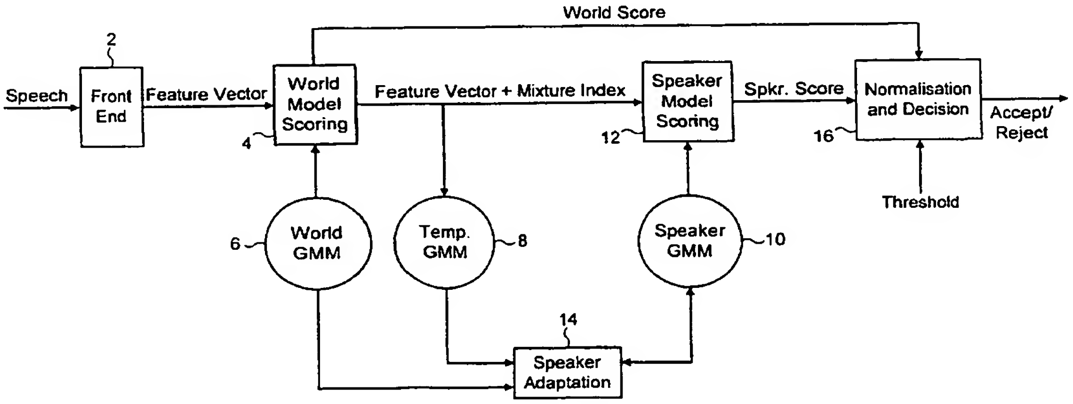
(81) **Designated States** *(national)*: JP, US.

(84) **Designated States** *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) **Title:** SPEAKER VERIFICATION



(57) **Abstract:** A speaker verification system is provided for identifying whether an input portion of speech originated from a particular speaker. A set of features is extracted from an input portion of speech provided by the speaker. A first scoring means (4) scores the set of features with a first stored model of mixture components derived from sets of features extracted from input portions of speech provided by a plurality of speakers. A second scoring means (12) scores the set of features with a second stored model of mixture components derived from sets of features extracted from input portions of speech provided by the speaker to be identified. The results are compared to determine whether the input portion of speech did originate from that particular speaker. The system provides that the first scoring means (4) scores the set of features with only part of the first stored model most likely to provide a good match to the set of features provided.

WO 02/067245 A1

SPEAKER VERIFICATION

This invention relates to a speaker verification system and in particular to a speaker verification system based on the principles proposed in our British patent application serial no. GB-A-2248513.

Speaker verification is important in applications such as financial transactions which are carried out automatically by telephone. Some of the problems of speaker verification are reduced by forming what are known as Gaussian Mixture models (GMM) for a number of utterances using features of these utterances from a large number of speakers. These models are known as world models. In addition, for every person whose speech is to be recognised, a GMM is formed. These models are known as personal or speaker models and comprise mixture components with which input utterances will be processed.

In speech verification, a person says an isolated or connected utterances and features from each of these utterances are extracted and into feature vectors. After this, the probabilities that these features vectors could have been generated for these words by the world model and by the personal model of that person are calculated and these probabilities are compared for the utterances. A decision on a verification for the speaker is then based on a poll of these comparisons.

More particularly, a system such as this operates by cutting an incoming stream of speech data into short sections or frames to allow feature extraction. A front end process extracts a set of features from each frame, these features being a function of the input speech signal. These features are then stored as a vector. The feature vectors are then further used for comparison to a world and speaker model.

During the verification process a sequence of feature vectors is processed with both the world and the speaker model. Only the features of the speaker model which best match the world model are processed and therefore only a small number of mixture components need to be processed with a speaker model. Therefore, most of the processing is needed for computing the world model.

For example, a GMM with 1024 mixture components would require each of these to be processed with an input vector and if a speaker model only had the best five mixture components then only these five would need to be processed with an input vector. Therefore, the world model processing needs considerably more of the computation in the verification system.

The present invention seeks to reduce the processing of the mixture components from the world model thereby considerably reducing the computational overhead of the whole process.

The invention is defined with more precision in the appended claims to which reference should now be made.

A preferred embodiment of the invention will now be described in detail, by way of example, with reference to the accompanying drawings in which:

FIGURE 1 shows a block diagram of a system embodying the invention;

FIGURE 2 shows a block diagram of a basic world model scoring system; and

FIGURE 3 shows a component predicted world model scoring system in accordance with an embodiment of the invention; and

FIGURE 4 shows schematically how component prediction is performed.

The diagram of Figure 1 shows an input speech signal to a front end processor 2 which produces as its output a feature vector. This is achieved as described above by cutting the speech signal into frames and from each frame extracting a set of features which are then combined into a feature vector for that frame.

The next stage in the process is that the feature vectors are provided to a world model storing unit 4. This also receives, as an input, mixture components from a world GMM which comprises mixture components for all possible speakers. The scoring process with the world model leads to a ranking of the mixture components according to the likelihood score for the given feature vector. It will be appreciated that the score for the comparison of each feature vector with each mixture component can only give a likelihood as there will be small variations in input speech every time a speaker provides an input signal. These variations will also occur as a result of the frames from which the feature vectors are extracted having cut points at different times in the speaker's speech each time speech is analysed. Therefore, there will be no exact matches between feature vectors and mixture components from the world model. All that will be produced is a likelihood score for the given feature vector corresponding to one of the world model mixture components.

After this likelihood scoring process has taken place, each feature vector is assigned to the best scoring mixture component of the world model and these are output from the world model score unit 4. The assigned feature vectors are accumulated in a temporary GMM model 8. This temporary model is used for a speaker adaptation process in conjunction with the world model to create a speaker model 10. The intention is to produce a speaker model which is a statistical representation of the speaker's speech, where each of the speaker's mixture components has exactly one corresponding component in the world model. This speaker model can then be used in a speaker model scoring unit 12.

A fast convergence of the speaker model parameter is one of the most important tasks in the system and is performed by the speaker adaptation unit 14. The speaker model parameters should change almost immediately to the input characteristic during the initialisation period. After a certain time, when enough speaker data are collected, the system should change from speaker adaptation to speaker tracking.

The tracking allows the system to follow changes in the speaker's voice pattern over a longer period of time. The tracking should be slow to allow capturing the voice over a long time span and not only over the last few utterances. This should lead to a more robust estimation of the speaker's model parameters.

The speaker adaptation unit 14 performs operations based on the standard equation for on-line model re-adaptation. In the case of background pre-whitened feature vectors, the equation is:

$$\hat{\mu}_{s_{n+m},r} = \underbrace{\frac{n}{n+m+8}}_{=\gamma}\mu_{s_n,r} + \underbrace{\frac{m+8}{n+m+8}}_{=1-\gamma}\hat{\mu}_{s_m,r}$$

where $\hat{\mu}_{s,\cdot r}$ is an already estimated speaker model parameter from the speaker model 10 which represents *m* seen frames, $\mu_{s,\cdot r}$ is the new mean accumulated over the last segment in the temporary model 8 with the representative weight of *n*. The new re-adapted model parameter $\hat{\mu}_{s,\cdot r}$ represents *n+m* frames and is calculated with weights, according to *n* and *m*, for $\hat{\mu}_{s,\cdot r}$ and $\mu_{s,\cdot r}$ respectively. The numeral 8 is only a preferred value, and the values may be appropriate in other circumstances.

A problem of the accumulation is the memory usage for each parameter. $\mu_{s,\cdot r}$ can be stored with 16 bit resolution whereas $\hat{\mu}_{s,\cdot r}$ is stored with only 8 bit. The averaging is not very accurate if the resolution for $\mu_{s,\cdot r}$ is reduced to 8 bit. Therefore the storage of the temporary model 8 is twice as large as the speaker model 10.

A simple way of reducing the memory size of the temporary model is to store only a sub-set of all mixture components in the temporary model. This enables the memory to be reduced to a half of the original size. The components are chosen by the frequency of their occurrence. Components with a high frequency are kept in memory in the temporary model. If a frame is seen for a component which is not in the memory, the component with the lowest frequency in the memory is checked and possibly exchanged with the new component from the world model 6.

The disadvantage of the temporary model reduction is the loss of information. Therefore the speaker model estimation will take longer.

Speaker tracking uses the same equation as in adaptation. The weighting factor γ is set to a certain value instead of being calculated individually according to the number of seen frames.

$$\hat{\mu}_{S_{n+m},r} = \gamma \mu_{S_n,r} + (1-\gamma)\hat{\mu}_{S_m,r}$$

The adaptation will change to a tracking approach if enough frames have been received to train the mixture component parameters for the target speaker. In this system, this is the case when 255 frames have been processed for a certain mixture component, but other values could be selected.

There are several ways to choose γ. The value should not be too large to avoid a fast tracking which weights the newer data very high and therefore loses information from the past quickly. If γ is chosen too low, the target speaker's change might not be captured and the speaker is locked out by the system. The value for γ also depends on the model size and the length of the test segment.

To validate performance of the system, first a test with standard adaptation is performed. This reveals the baseline performance for further testing of memory reductions on the temporary model and the speaker tracking settings. Tests are performed with a model size of 64 mixture components. Larger model sizes might not obtain any changes between the different tracking strategies due to the small amount of training sessions available. A comparison is also made for gender dependent background models and a combined gender model for the adaptation process.

Once the mixture components have been inserted in the speaker model, the feature vectors and mixture components output by the world model scoring unit are provided to the speaker model scoring unit 12. Producing a likelihood score for the speaker model involves only the processing of the most likely mixture components with the feature vectors. These components also retain high scores from the speaker model due to the component correspondence between speaker and world model. Therefore, only a small number of components are processed for scoring the speaker model.

The output scores of the world and speaker model scoring units are likelihood scores which are input to a normalisation and decision unit 16. The two world and speaker model scores are normalised by subtracting the world model score. This is compared to a threshold and the speaker is accepted or rejected by the system in dependence on the difference signal or falling above or below the threshold. An accept or reject signal is then output by the normalisation and decision unit 16.

A process known as component prediction can be used to speed up the processing of the world model scoring. It will be appreciated that in the diagram described above, each of the input feature vectors from the front end processor 2 has to be compared with each of the mixture components from the world model 6 in the world model scoring unit 4. This task is therefore computationally very expensive both in initial training of the unit for a speaker and for subsequent testing.

To understand how component prediction works, the standard world model scoring system is first explained in more detail without component prediction. This is done with reference to Figure 2 which shows the task of world model scoring.

The world model, which is a GMM, consists of a number of mixture components 20.

In the scoring process, each of the mixture components 20 in the world model is processed with an input feature vector in a scoring unit 22. The result of this scoring is a likelihood score for each of the mixture components. The likelihood of the scores for all of the components are sorted according to their values to produce a likelihood ranking of the mixture components. The best scoring components recognised by their indices are used for further processing and are stored for each feature vector in a best scoring component store 24.

The other output from the scoring unit 22 is the world score. For each feature vector, this is calculated by combining the likelihood scores of the best scoring component.

In Figure 3, a component predicted world model scoring system is shown which illustrates the extension of the world model scoring using component prediction.

In this, the indices of the best scoring components stored at 24 are used to choose indices from a look-up table index 26. These indices point to information about which mixture components are most likely to obtain high likelihood scores for the following feature vector. Thus, these contain data about which mixture components should be selected for comparison with the next input feature vector. Only a subset of all the components is selected, eg. 5 components according to the data from the look-up table. This component selection is performed by component selection unit 28 before being provided to the scoring unit 22 for scoring with the next feature vector. The scoring of this next feature vector again leads to best scoring components and the indices of these components are again used for a prediction for the following feature vector. Thus, only a small number of mixture components from the world model are processed with each feature vector thereby considerably reducing the computational overhead. The saving will depend on the exact number of mixture components selected.

The look-up tables which are referenced by the indices of best scoring components are now described in more detail.

Given the sequence of feature vectors, which is a known sequence, the idea is to predict certain mixture components from the world model which are most likely to achieve high scores for processing of the immediately succeeding vector. This prediction is based on a data driven estimation of the most likely component indices. For example, a total of 25 components might be predicted from a total of 1124 mixture components in a world model thereby reducing the processing time by over 95%.

The prediction is derived from transition probabilities for transversing from a mixture component J to a mixture component I in acoustic space.

The Gaussian component of a world GMM can be trained using the EM-algorithm published in the Statistical Society, 39:1-38,1997 by a Dempster, A., Laird, N., and D. Rubin under the heading "Maximum likelihood from incomplete data via the EM algorithm". This algorithm assumes equal probabilities for all state transitions. However, transition probabilities can be calculated after training of the mixture components. These transition probabilities allow prediction of certain mixture components for the processing of the succeeding vector.

Figure 4 shows an overview of the prediction scheme. A feature vector is processed using the world GMM. A likelihood score of the feature vector is calculated to each mixture component (Stage 1). These scores are sorted. The look-up tables of the most likely mixture components are used for prediction (stage 2). The component indices of these tables are copied into a component prediction array for the processing of the next feature vector.

Several parameters can be varied in the prediction process. These are the number of mixture components used for prediction and the size of the look-up tables. Another aspect is the processing of the first feature vector in a vector sequence. The prediction may not be used for the first frame which is similar to a calculation of all components in the GMM.

Most of the times the prediction produces the same substantially same result as full processing. Sometimes the prediction deteriorates with the frame likelihood score but the prediction does not degenerate into a random component calculation. The initial component prediction obtains good results when used for the first frame of a speech segment. This is only in general, the differences between the two initialisation methods are minor at the start of a speech segment.

Global transition estimates can be averaged over all consecutive vector pairs. An initial transition probability for the start of a vector sequence can be defined for a number of extracted speech segments.

The transition probabilities P(I:J) are sorted for each component J and only the most likely indices of I are stored in the look up table 26. The table for each mixture component can vary in size. When a feature vector is processed with the world GMM it leads to a most likely mixture component. This is stored in the indices of aposteirori components 24 and is used to select the look up table to be used by the component selection unit 28. The table contains the indices of the mixture components which will be processed for the next feature vector, these mixture components being the most likely mixture components to follow the current feature vector.

Returning now to Fig. 1, it will be explained how component prediction improves the performance of the system of Figure 1 for both the initial training of the speaker model and for subsequent testing of a speaker.

When a new speaker model is to be created for a new user this is done by the speaker saying a known sequence of words a certain number of times and each utterance of the word being used to generate the most likely components from the world model which correspond to that speaker. This is done by first using the world model scoring unit 4 with extracted feature vectors. Initially, the first feature vector is scored with all the components of the world model 6 and an index of the best scoring components with this stored at 24. The look up table store 26 then provides data corresponding to the most likely set of components which should be compared with the next feature vector. This most likely set is the most likely set of the world components not of any particular speaker's components. These are then scored with the next input vector and the process repeats.

At the same time, the temporary model 8 receives the feature vectors output by the world model scoring unit 4 and a speaker adaptation unit 14 uses this to produce a speaker model 10 for that particular speaker. Thus, a speaker model is created and is stored for future reference.

In testing, a speech input is tested against a speaker model. This can be done by a user who is to be identified first inputting, for example, a first identification number or some other identifier. This causes the system to load what it believes to be a speaker model for that speaker into the speaker model store 10. An utterance of speech from the speaker is then processed by the front end processor and the world model scoring unit which operates according to the system Figure 3. That is to say, it first scores the first input vector with all the components from the world model 6 before using the index of best scoring components 24 and look up table store 26 to select by component selection unit 28 the most likely set of components for scoring against the next vector.

Feature vector mixture indices are supplied to the speaker model scoring which operates only with the mixture components in the speaker model 10. Thus, the processing of vectors is significantly reduced and thereby the computational overhead. It will be appreciated that in training the full potential of the speed up in computation is achieved due to the processing of just one speaker model.

CLAIMS

1. A speech verification system for identifying whether an input portion of speech originated from a particular speaker comprising:

means to extract a set of features from an input portion of speech provided by the speaker;

first means for scoring the set of features with a first stored model of mixture components derived from sets of features extracted from input portions of speech provided by a plurality of speakers;

second means for scoring the set of features with a second stored model of mixture components derived from sets of features extracted from input portions of speech provided by the speaker to be identified;

means for comparing results provided by the first and second scoring to determine whether the input portion of speech did originate from the said particular speaker;

characterised in that

the first scoring means scores the set of features with only the part of the first stored model most likely to provide a good match to the set of features.

2. A speech verification system according to claim 1 in which the first scoring means includes indices of best scoring mixture components for sets of features, a look up table containing data identifying the portions of the first stored model most likely to provide a good score with the next set of features, and means for selecting the said portions of the first stored model for scoring with the next set of features.

3. A speech verification system according to claims 1 and 2 in which the extracting means extracts a plurality

of sets of features from the input portion of speech and each set of features comprises feature vectors derived from the sequence of sounds in an interval.

4. A system according to claim 1 in which the system is arranged to initialise to a new speaker during an initialisation period, wherein the second means for scoring the set of features does this a predetermined number of times for the new speaker scores them with an estimated set of mixture components in the second stored model and modifies the estimated set of mixture components after each iteration of the scoring means.

5. A system according to claim 1 or 4 in which the second stored model is modified if the comparison means indicates that the input speech came from a particular speaker and the stored model is modified in dependence on the result of the scoring of the set of features with the second stored model.

6. A system according to claim 5 in which a weighting factor is used to modify the second stored model with the set of features.

7. A method for speech verification for identifying whether an input portion of speech originated from a particular speaker comprising the steps of:

extracting a set of features from an input portion of speech provided by the speaker;

scoring the set of features with a first stored model of mixture components derived from sets of features extracted for input portions of speech provided by a plurality of speakers;

second means for scoring the set of features with a second stored model of mixture components derived from sets of features extracted from input portions of speech provided by the speaker to be identified;

comparing the results provided by the respective scoring steps to determine whether the input portion of speech do originate from the speaker to be identified;

characterised in that the scoring of the set of features with the first stored model of mixture components comprises scoring the features only with the part of the first scored model most likely to provide a good match to the input set of features.

8. A method according to claim 4 in which the part of the first stored model most likely to provide a good match to the set of features comprises looking up previously stored data corresponding to the most likely set of components which should be compared with the next feature vector from the first scored model and scoring this set of mixture components with the next set of features.

9. A method according to claim 7 including the step of initialising to a new speaker during an initialisation period in which the scoring of the set of features with the second stored model is repeated a predetermined number of times commencing with an estimated set of mixture components in the second stored model, and modifying the estimated set of mixture components with each iteration.

10. A method according to claim 7 or 9 including the step of modifying the second stored model if the company step indicates that the input speech come from a particular speaker in dependence on the result of the scoring of the set of features with the second stored model.

11. A method according to claim 10 in which a weighting factor is used in the modifying step.

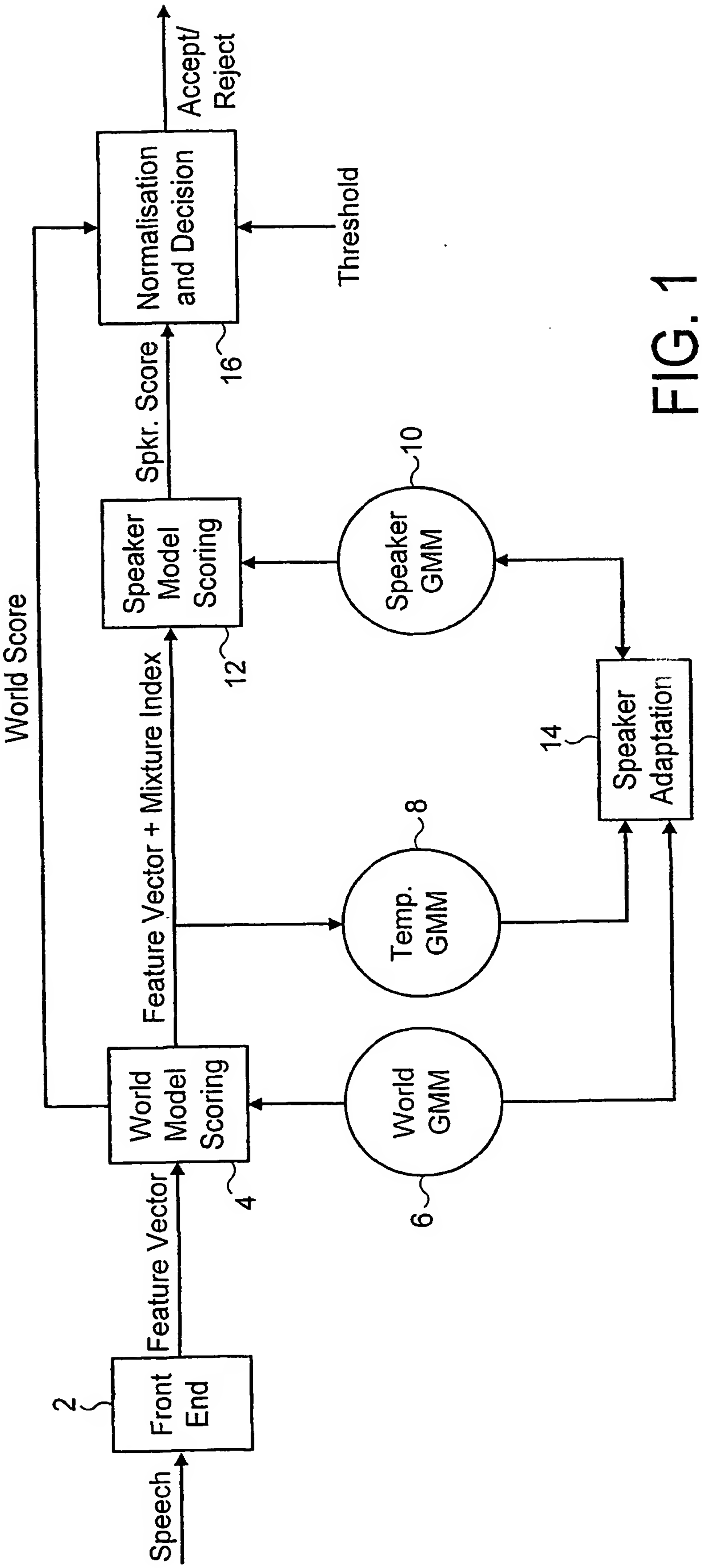

Speech
Front End
2
Feature Vector
World Model Scoring
4
World GMM
6
World Score
Feature Vector + Mixture Index
Temp. GMM
8
Speaker Model Scoring
12
Speaker GMM
10
Speaker Adaptation
14
Spkr. Score
Normalisation and Decision
16
Threshold
Accept/ Reject

FIG. 1

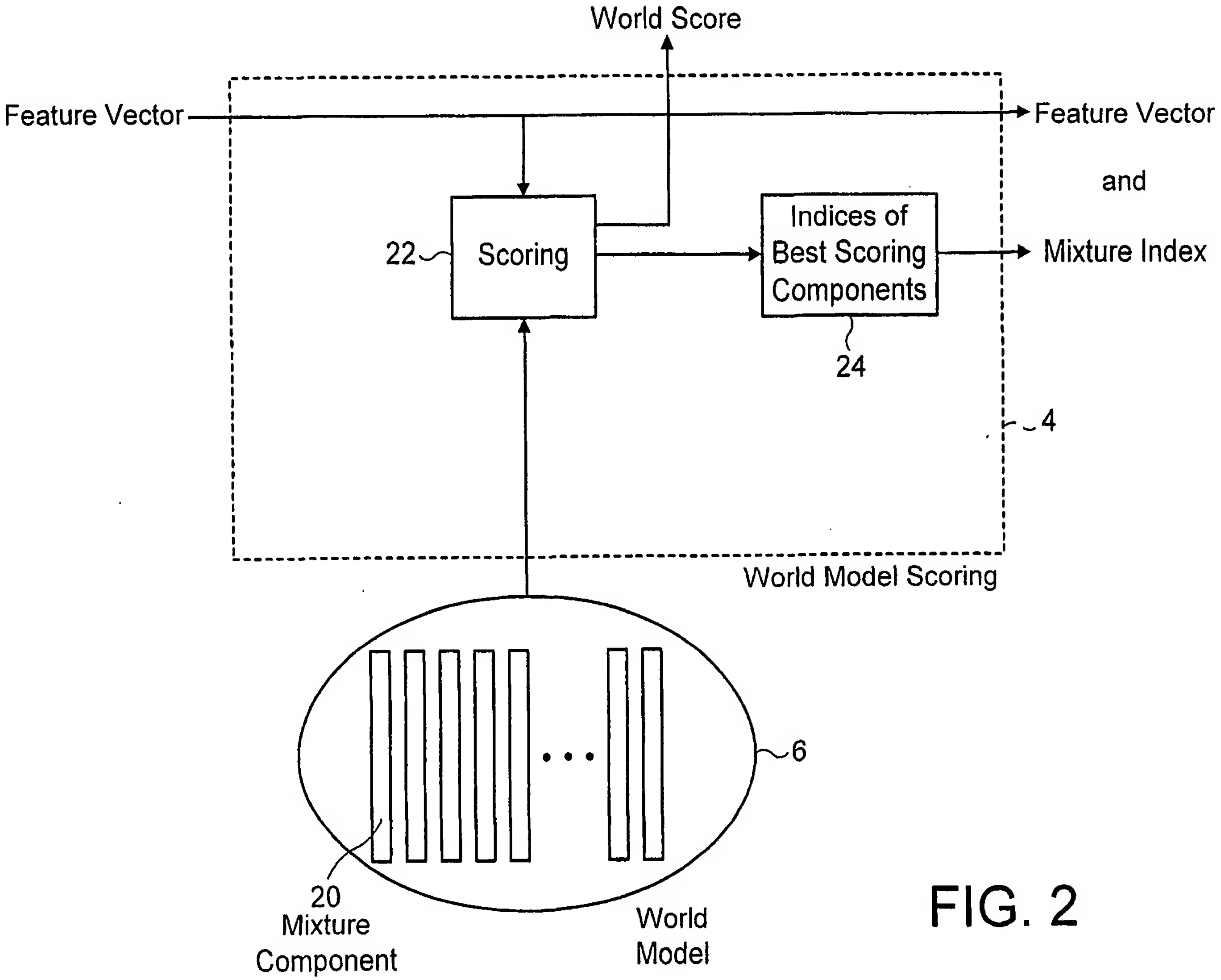

World Score
Feature Vector
Feature Vector
and
Mixture Index
22
Scoring
Indices of
Best Scoring
Components
24
4
World Model Scoring
6
20
Mixture
Component
World
Model

FIG. 2

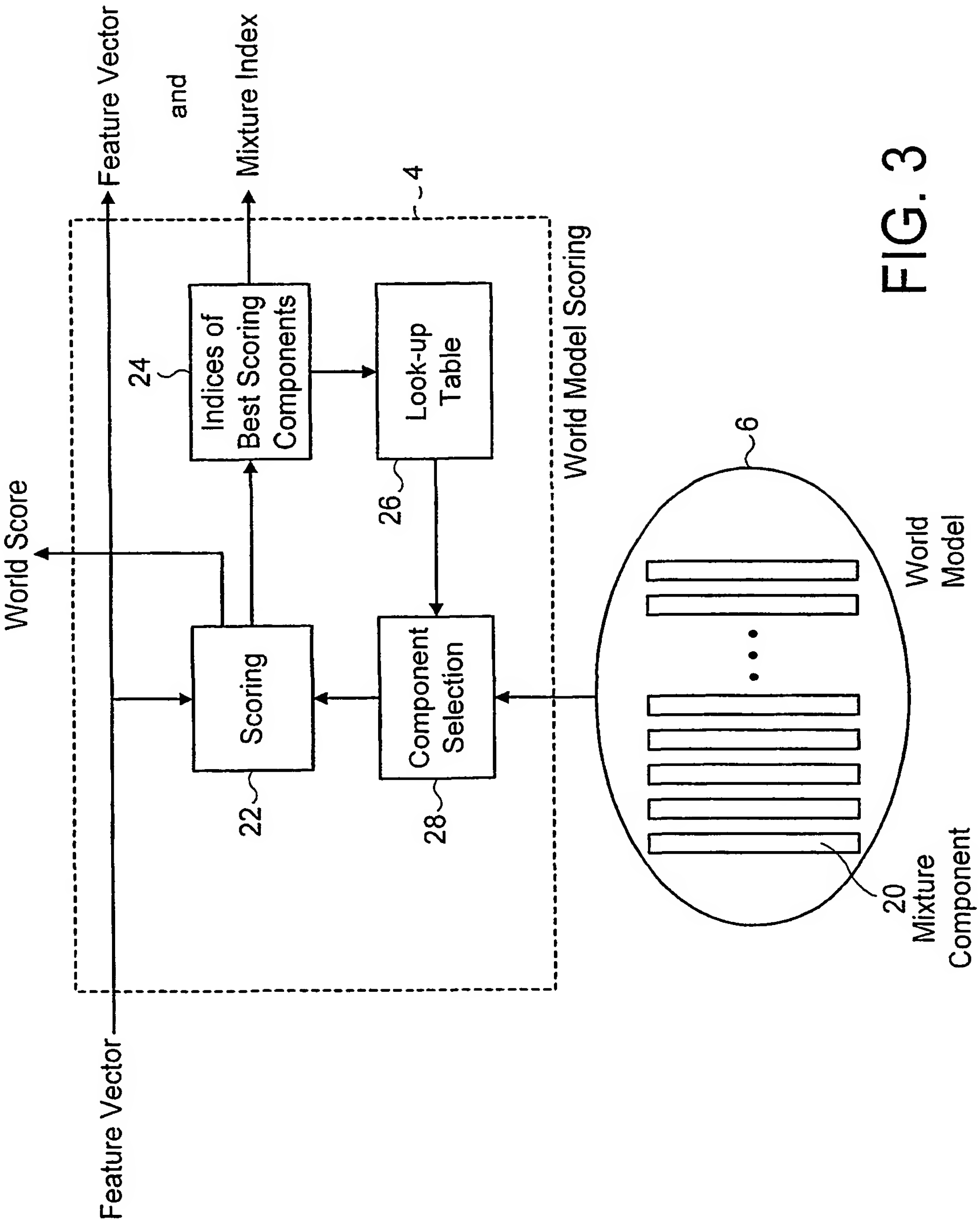

Feature Vector
Feature Vector
and
Mixture Index
World Score
4
World Model Scoring
24
Indices of Best Scoring Components
Look-up Table
26
Scoring
22
Component Selection
28
6
World Model
20
Mixture Component

FIG. 3

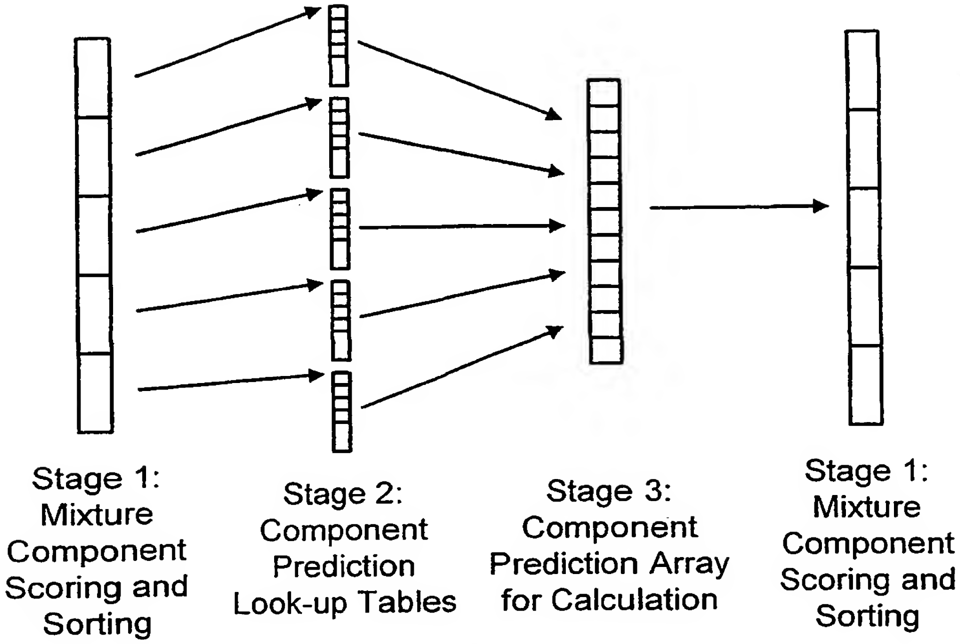

Stage 1:
Mixture
Component
Scoring and
Sorting
Stage 2:
Component
Prediction
Look-up Tables
Stage 3:
Component
Prediction Array
for Calculation
Stage 1:
Mixture
Component
Scoring and
Sorting

FIG. 4

INTERNATIONAL SEARCH REPORT

Inter ıal Application No
PCT/GB 02/00665

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L17/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, INSPEC, WPI Data, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	REYNOLDS D A: "COMPARISON OF BACKGROUND NORMALIZATION METHODS FOR TEXT-INDEPENDENTSPEAKER VERIFICATION" 5TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. EUROSPEECH '97. RHODES, GREECE, SEPT. 22 - 25, 1997, EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. (EUROSPEECH), GRENOBLE: ESCA, FR, vol. 2 OF 5, 22 September 1997 (1997-09-22), pages 963-966, XP001004029 abstract page 963, right-hand column -page 964, right-hand column --- -/--	1-11

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search 14 May 2002	Date of mailing of the international search report 24/05/2002
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Ramos Sánchez, U

INTERNATIONAL SEARCH REPORT

Intei ial Application No
PCT/GB 02/00665

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	NAKAGAWA S ET AL: "Speaker verification using frame and utterance level likelihood normalization" ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1997. ICASSP-97., 1997 IEEE INTERNATIONAL CONFERENCE ON MUNICH, GERMANY 21-24 APRIL 1997, LOS ALAMITOS, CA, USA,IEEE COMPUT. SOC, US, 21 April 1997 (1997-04-21), pages 1087-1090, XP010225987 ISBN: 0-8186-7919-0 page 1087, right-hand column -page 1088, right-hand column ---	1-11
A	EP 0 822 539 A (DIGITAL EQUIPMENT CORP) 4 February 1998 (1998-02-04) page 4, line 24 - line 49 ---	1-11
A	GB 2 248 513 A (ENSIGMA LTD ;SECR DEFENCE (GB)) 8 April 1992 (1992-04-08) cited in the application page 18, line 13 - line 35 -----	1-11

INTERNATIONAL SEARCH REPORT

Interr al Application No
PCT/GB 02/00665

Patent document cited in search report		Publication date	Patent family member(s)			Publication date
EP 0822539	A	04-02-1998	US	6205424	B1	20-03-2001
			EP	0822539	A2	04-02-1998
			JP	10083194	A	31-03-1998
GB 2248513	A	08-04-1992	AU	665745	B2	18-01-1996
			AU	8649691	A	28-04-1992
			WO	9206468	A1	16-04-1992
			US	5526465	A	11-06-1996
			ZA	9107886	A	28-10-1992